

Hard and soft bounds in the evolution of Ubuntu packages. A lesson for species body masses?

Marco Gherardi^{*1,2}, Salvatore Mandrà¹, Bruno Bassetti^{1,2}, and
Marco Cosentino Lagomarsino^{3,4}

¹*Dipartimento di Fisica, Università degli Studi di Milano, via Celoria 16, 20133 Milano, Italy*

²*Istituto Nazionale di Fisica Nucleare, Sezione di Milano, via Celoria 16, 20133 Milano, Italy*

³*Genomic Physics Group, UMR 7238 associée au Centre National de la Recherche Scientifique,
“Microorganism Genomics,” 15 rue de l’École de Médecine, 75006 Paris, France*

⁴*Université Pierre et Marie Curie, 4 Place Jussieu, 75005 Paris, France*

March 4, 2013

Understanding the patterns of software evolution has a large practical importance: the knowledge of what can be considered “typical” can guide developers and engineers in recognizing and reacting to abnormal behavior. While the initial framework of a theory of software exists [1, 2], the current theoretical achievements do not fully capture existing quantitative data or predict future trends. Programs are embedded in the real world, and consequently the growth of a software package is characterized by inherent adaptive change in response to complex factors. The multi-level feedback where programs and their environment evolve in concert is elusive and difficult to describe precisely.

These very features make the subject attractive from the point of view of “complex systems” theory and analysis. Most of the “traditional” analyses concerned proprietary software, but a number of studies carried out within the past 10-15 years gathered a relevant amount of evidence concerning the evolution of Open Source Software (OSS) [3, 4, 5, 6]. The open source phenomenon has two specificities that make it particularly interesting. First, the goal of an open source project is to create a system that is useful or interesting to its developers, and thus fills a “social void” rather than a commercial one. Second, large OSS projects are developed and maintained in a globally decentralized context, contrary to traditional software.

^{*}Marco.Gherardi@mi.infn.it

The emergent complex self-organizing structure challenges traditional theories of management and engineering [7, 8]. The OSS phenomenon is also affecting the daily lives of increasingly many people, since OSS operating systems and applications run on devices ranging from PCs to mobile phones and tablets.

Perhaps the simplest observable related to software growth is its size, which can be measured with different approaches [9]. Despite its simplicity, the size of a piece of software encapsulates many of the features of its evolution and evolvability. Here, we consider the dynamics of package size in a widely used GNU/Linux system, the Debian-based Ubuntu distribution (www.ubuntu.com/project). We analyze systematically the available data and show that they are compatible with a multiplicative anomalous diffusion process. We study this process with the aid of a theoretical model, and show that the combination of a “hard” lower cutoff and a “soft” upper cutoff on package size reproduces with extreme accuracy the observed distribution. The same model makes definite quantitative predictions for the *future* dynamics of Ubuntu packages. Finally, as we will see, the knowledge of these evolutionary patterns might lend a fresh perspective to the debate on the quantitative aspects of an *a priori* unrelated process, the cladogenesis that determines the mass distribution of mammals.

Ubuntu “packages” are bundled files comprising the pieces of software that make up the whole system. Since Ubuntu was first released in October 2004, the number of packages increased from a few hundred to tens of thousands. Since then, one new release every six months has been issued. This chronological regularity is valuable for a systematic quantitative study. The first, second, and third releases were christened *Warty Warthog*, *Hoary Hedgehog*, and *Breezy Badger*; from then on, the naming followed alphabetical order, encompassing 17 different real and imaginary animals, up to the latest *Quantal Quetzal* (October 2012).

Analysis of empirical data for approximately 370 000 changes in package size, between successive releases spanning the entire lifetime of Ubuntu, reveals striking regularity (Fig. 1, top panel). The logarithm of the multiplicative change $\delta = s'/s$ between the sizes s and s' of a package in consecutive releases appears to follow an “ α -stable” distribution, independently of the initial size s and of time (the distribution is centered in $\delta = 1$ and has power-law exponent $\alpha \approx 0.7$). This class of distributions, widely used in many modeling contexts [10, 11, 12, 13], contains the most general probability distributions followed by the sum of a large number of independent identically-distributed random variables; it is therefore a generalization of the Gaussian (which is recovered for $\alpha = 2$). Such a multiplicative growth is reminiscent of the trends found in another area of human interactions, the evolution of business firms’ sizes [14].

Notably, while the bulk of the empirical distribution of multiplicative size changes is symmetric, events belonging to the tails are bounded in a size-dependent way (Fig. 1, bottom panel). No package can shrink to sizes smaller than a global

cutoff s_{\min} . This *hard* bound is easily rationalized by the existence of minimum requirements from the package management system. Consequently, the largest possible decrease is $\delta_{\min} = s_{\min}/s$; from the empirical data we fix $s_{\min} = 741$ bytes, which is the average of the minimum package sizes found in each release. Such small packages are usually “dummy” packages, not containing any functional software, but typically pointers to other packages. Expansion to larger package sizes manifests a more intriguing and complex behavior: the largest size that a package can attain between two consecutive releases *depends on its starting size*. Specifically, the largest possible increase is $\delta_{\max} = (s_{\max}/s)^\gamma$, with an exponent γ approximately equal to $1/2$. We call this a *soft* bound, meaning that larger packages are less prone to perform large jumps, but packages of different initial sizes do not behave as if a unique maximal size were present. Trade-offs between increase in complexity and cost of deployment are probably responsible for this non-linear law. Birth of new packages is also a relevant driving process for the dynamics of software evolution. Newborn packages appear with size proportions approximately equal to those of the preceding release, which suggests that the dominant route of expansion is software forking and reuse (but note that the data set does not distinguish new packages from old packages with new names).

Based on the foregoing empirical observations, we define a stochastic model of package size evolution, which relies on three assumptions: (i) At every new release, each package (of size s) assumes the new size $s' = s\delta$ (multiplicative size changes). (ii) Each package has a small probability p of also adding a copy of itself to the new release (branching). (iii) The logarithms of the growth factors δ are independent α -stable random variables conditioned on two size-dependent cutoffs, a lower hard bound and an upper soft bound, as evidenced by the data. Technically, this model is realized as a branching multiplicative diffusion process. We do not explicitly consider package deletion, which does happen in Ubuntu; however, we found that its role is irrelevant for the evolution of package size distributions. The model above has no free parameters, as all the quantities needed to specify the distribution are estimated by data analysis.

Starting from the population of packages in the first Ubuntu release, *Warty*, and evolving their sizes for 16 steps (eight years), the model predicts very accurately the package size distribution in the latest release, *Quantal* (Fig. 2). Sensitivity analysis shows that the results are robust with respect to variation of the parameters. Moreover, the accordance of model and data is not dependent on the particular shape of the distribution; in fact, arbitrarily chosen subsets of packages can be followed through their evolution, and the size proportions they assume in *Quantal* are reproduced strikingly well. The model is predictive, and can be used to forecast future evolution. For instance, we found that the current distribution is very far from stationary; at this rate, a stationary state would be reached in approximately 2–400 years. In 10 years the largest package should weigh approximately 1 Gb,

and the average package size is predicted to nearly double from the current 1.2 Mb to about 2.3 Mb; the most common size, instead, will have slightly increased only by around 10 kb (it is currently 22 kb).

We found that the knowledge of the anomalous diffusion framework with “soft” cutoffs described above may suggest a different perspective on the debate around a distant scientific problem. In fact, similar models to the one described here have been employed to explain the evolution of species body masses in mammals and other taxa [15, 16]. In this case, the branching process represents cladogenesis, i.e. the lineage splitting event generating new species (*clades* in the phylogenetic tree) whose average body mass is related to the ancestor’s. The model proposed by Clauset and Erwin [16] and further developed in [17, 18] assumes multiplicative diffusion on evolutionary time scales, with a lower hard bound due to metabolic constraints, and an explicit bias toward larger sizes (the controversial “Cope’s rule” [19, 20, 21, 22]), whose strength must increase for lower masses (although there appears to be also evidence for the opposite tendency [15]). Moreover, the introduction of a size-dependent extinction rate is necessary in order to approximate the large-mass tail of the empirical distribution of extant mammals.

In the framework suggested by software evolution, it seems natural to characterize the low propensity of large species to generate exceedingly large descendant species (and the tendency of small species to generate larger ones) through a “soft” cutoff instead. Fossil data of ancestor-descendant size ratios are not abundant, and susceptible to noise and bias [23]. We used a compilation by Alroy of 1109 North American terrestrial mammals up to the late Pleistocene, obtained by a highly conservative method [15]. Despite the great amount of work behind these data, they do not allow an estimate of parameters nearly as precise as what was attained for Ubuntu packages; nonetheless, our analysis shows that the changes in body size are compatible with upper and lower soft cutoffs with γ -values around 0.2 and 0.6 respectively (see Fig. 3, top panel). Uncertainties on these estimates are not a big inconvenience, as the results are fairly robust to variation of these parameters.

We simulated the *in silico* evolution of body masses throughout mammalian history, starting from the body mass of the founder species *Hadrocodium wui*, a small mammaliaform from the Early Jurassic weighing 2 grams [24]. Remarkably, the characteristically skew and wide distribution of extant mammals [25] is recovered with good precision by this model (Fig. 3, bottom panel). The (softly) bounded nature of the diffusion, together with the asymmetry of the initial condition, are the key ingredients that account for the shape of the empirical distribution. It must be said that the agreement is not completely parameter-free as in the case of Ubuntu packages: model time is chosen as the one that best recovers the expected distribution, since it cannot be estimated directly. However, one or more free parameters were present also in all previous studies [16, 17].

One important remark is that the present model relaxes the common assumption

that the mammalian body-mass distribution is stationary at present time. Consequently, different initial conditions can produce markedly different distributions. If the initial mass is sufficiently large, left-skewed distributions can be obtained; such a shape is less common but it is nonetheless found in some taxa [26]. Mining the literature, we could not find any strong evidence either supporting or undermining the assumption of stationarity, and therefore we hope that our findings could be useful to stimulate the debate in this direction. Incidentally, we note that a further prediction of the model is a slowly saturating evolution for the maximum body mass as a function of time, which is in line with recent findings [27].

A second and final remark is that the bounds on the diffusion process in the context of mammalian body masses are realized by a size-dependent extinction rate [23]. In our approach, the soft nature of the constraint for large masses is interpreted as the result of the competition between the short-term selective advantages of an increased body size and the corresponding long-term extinction risk, as concluded in previous studies. This macroevolutionary tradeoff mechanism is quantitatively robust across all mammalian species [28], and also in other taxa [18]. The observation that the lower boundary is soft as well suggests that a similar tradeoff may be present also for small body masses.

Acknowledgements We are grateful to Aaron Clauset for help with the Alroy dataset. We wish to thank Alberto Vailati, Vincenzo Gino Benza, and Matteo Osella for helpful suggestions.

References

- [1] M.M. Lehman and J.F. Ramil. Software evolution — background, theory, practice. *Inf. Process. Lett.*, 88(1-2):33–44, 2003.
- [2] T. Mens and S. Demeyer, editors. *Software Evolution*. Springer, 2008.
- [3] J. Fernández-Ramil, A. Lozano, M. Wermelinger, and A. Capiluppi. Empirical studies of open source evolution. In *Software Evolution*, pages 263–288. 2008.
- [4] L. Ermann, A. D. Chepelianskii, and D. L. Shepelyansky. Fractal Weyl law for Linux kernel architecture. *Eur. Phys. J. B*, 79:115–120, 2011.
- [5] T. Maillart, D. Sornette, S. Spaeth, and G. von Krogh. Empirical tests of Zipf’s law mechanism in open source Linux distribution. *Phys Rev Lett*, 101(21):218701, Nov 2008.

- [6] M.W. Godfrey and Q. Tu. Evolution in open source software: A case study. In *ICSM*, pages 131–142, 2000.
- [7] G. Madey, V. Freeh, and R. Tynan. *The open source software development phenomenon: An analysis based on social network theory*, page 1806–1813. 2002.
- [8] M.A. Fortuna, J.A. Bonachela, and S.A. Levin. Evolution of a modular software network. *Proceedings of the National Academy of Sciences*, 2011.
- [9] C.F. Kemerer and S. Slaughter. An empirical approach to studying software evolution. *IEEE Trans. Software Eng.*, 25(4):493–509, 1999.
- [10] B.B. Mandelbrot. *The Fractal Geometry of Nature*. Henry Holt and Company, 1982.
- [11] R.N. Mantegna and H.E. Stanley. Scaling behaviour in the dynamics of an economic index. *Nature*, 376:46, 1995.
- [12] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439:462, 2006.
- [13] P. Barthelemy, J. Bortolotti, and D.S. Wiersma. A Lévy flight for light. *Nature*, 453:495–498, 2008.
- [14] D. Fu, F. Pammolli, S.V. Buldyrev, M. Riccaboni, K. Matia, K. Yamasaki, and H.E. Stanley. The growth of business firms: theoretical framework and empirical evidence. *Proceedings of the National Academy of Sciences*, 102(52):18801–18806, 2005.
- [15] J. Alroy. Cope’s rule and the dynamics of body mass evolution in north american fossil mammals. *Science*, 280(5364):731–734, May 1998.
- [16] A. Clauset and D.H. Erwin. The evolution and distribution of species body size. *Science*, 321(5887):399–401, Jul 2008.
- [17] A. Clauset and S. Redner. Evolutionary model of species body mass diversification. *Phys. Rev. Lett.*, 102:038103, 2009.
- [18] A. Clauset, D.J. Schwab, and S. Redner. How many species have mass M ? *American Naturalist*, 173:256–263, 2009.
- [19] E.D. Cope. *The origin of the fittest*. Appleton, New York, 1887.
- [20] S.J. Gould. Cope’s rule as psychological artefact. *Nature*, 385(6613):199–200, 1997.

- [21] B. Van Valkenburgh, X. Wang, and J. Damuth. Cope's rule, hypercarnivory, and extinction in North American canids. *Science*, 306:101, 2004.
- [22] D.S. Moen. Cope's rule in cryptodiran turtles: do the body sizes of extant species reflect a trend of phyletic size increase? *Journal of Evolutionary Biology*, 19(4):1210–1221, 2006.
- [23] L.H. Liow, M. Fortelius, E. Bingham, K. Lintulaakso, H. Mannila, L. Flynn, and N.C. Stenseth. Higher origination and extinction rates in larger mammals. *Proceedings of The National Academy of Sciences*, 105:6097–6102, 2008.
- [24] Z.-X. Luo, A.W. Crompton, and A.-L. Sun. A new mammaliaform from the early Jurassic and evolution of mammalian characteristics. *Science*, 292:1535, 2001.
- [25] F.A. Smith, S.K. Lyons, S.K.M. Ernest, K.E. Jones, D.M. Kaufman, T. Dayan, P.A. Marquet, J.H. Brown, and J.P. Haskell. Body mass of late Quaternary mammals. *Ecology*, 84:3402, 2003.
- [26] J. Kozłowski and A.T. Gawelczyk. Why are species' body size distributions usually skewed to the right? *Functional Ecology*, 16:419–432, 2002.
- [27] F.A. Smith, A.G. Boyer, J.H. Brown, D.P. Costa, T. Dayan, S.K.M. Ernest, A.R. Evans, M. Fortelius, J.L. Gittleman, M.J. Hamilton, E. Larisa Harding, K. Lintulaakso, S.K. Lyons, C. McCain, J.G. Okie, J.J. Saarinen, R.M. Sibly, P.R. Stephens, J. Theodor, and M.D. Uhen. The evolution of maximum body size of terrestrial mammals. *Science*, 330(6008):1216–1219, 2010.
- [28] A. Clauset. How large should whales be? *PLOS One*, 8:e53967, 2013.

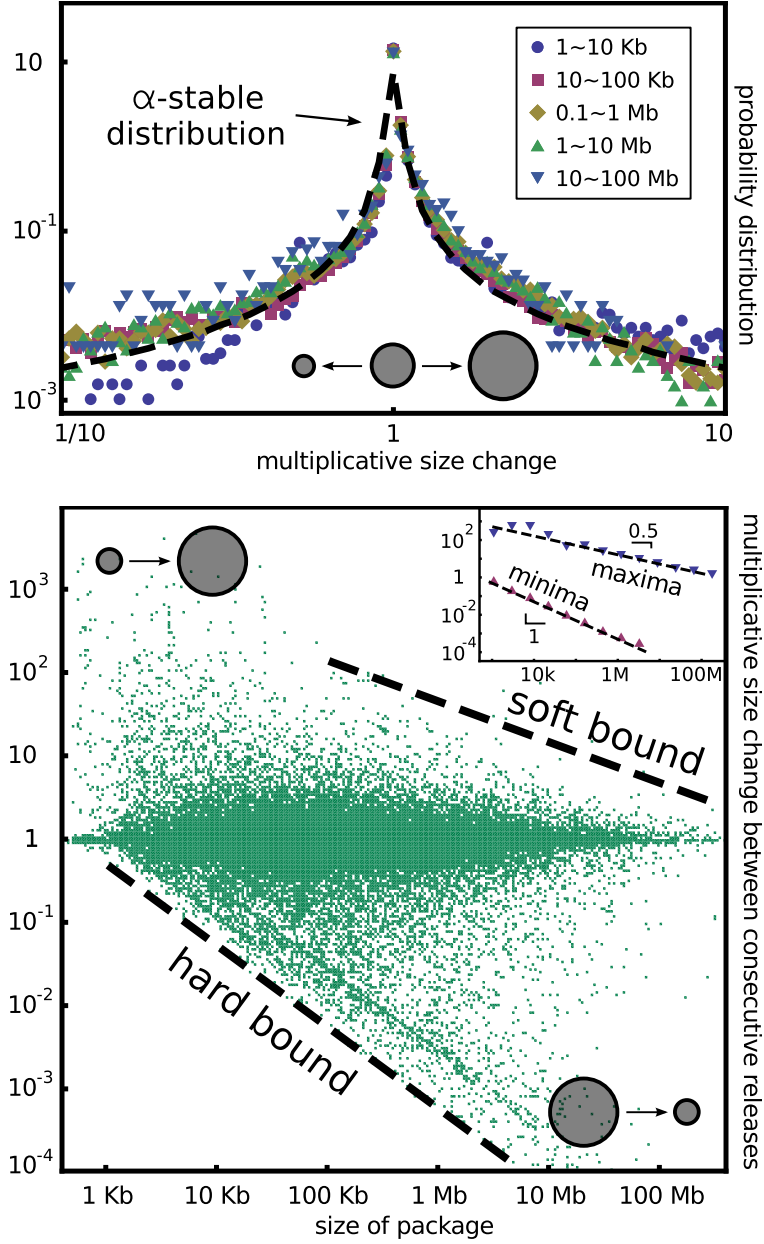


Figure 1: **The changes in package size between Ubuntu releases follow an α -stable distribution with size-dependent bounds.** (Top panel) The distribution is independent of size for small multiplicative size changes (x axis; symbols represent different size ranges); (bottom panel) a scatterplot of multiplicative size change vs initial package size in the whole range reveals a hard lower bound, due to a minimum attainable size, and a soft upper bound, with a nontrivial size dependence; the inset shows binned averages of maximum and minimum size changes (dashed lines are power-law fits yielding exponents 0.5 and 1 respectively).

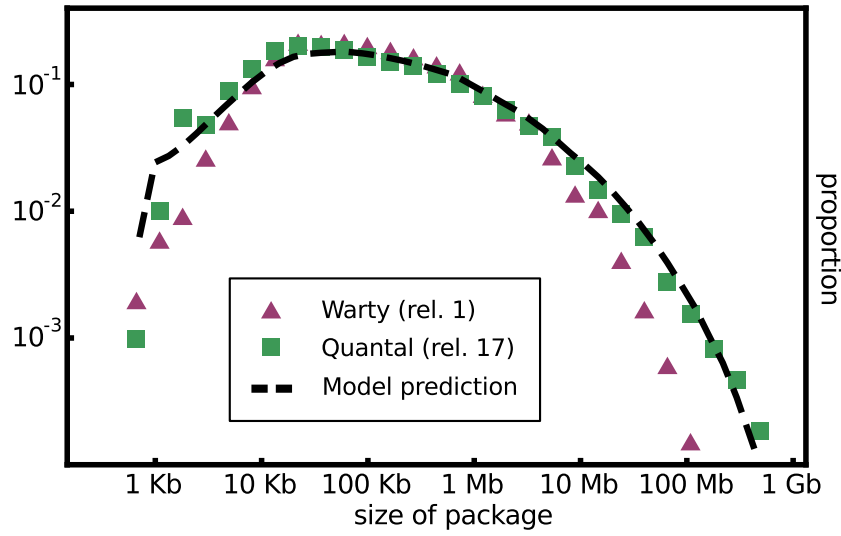


Figure 2: **The dynamics of package-size distribution is captured by a branching multiplicative diffusion process.** Starting from the initial pool of packages constituting *Warty Warthog* (triangles), with all parameters fixed by data analysis, the model yields the distribution traced by the dashed line, which nicely reproduces size proportions in *Quantal Quetzal* (squares). Notice that the tails of the two empirical distributions differ by almost one order of magnitude; furthermore, the ramp at small sizes for *Quantal* (which was not present in *Warty*) is correctly predicted.

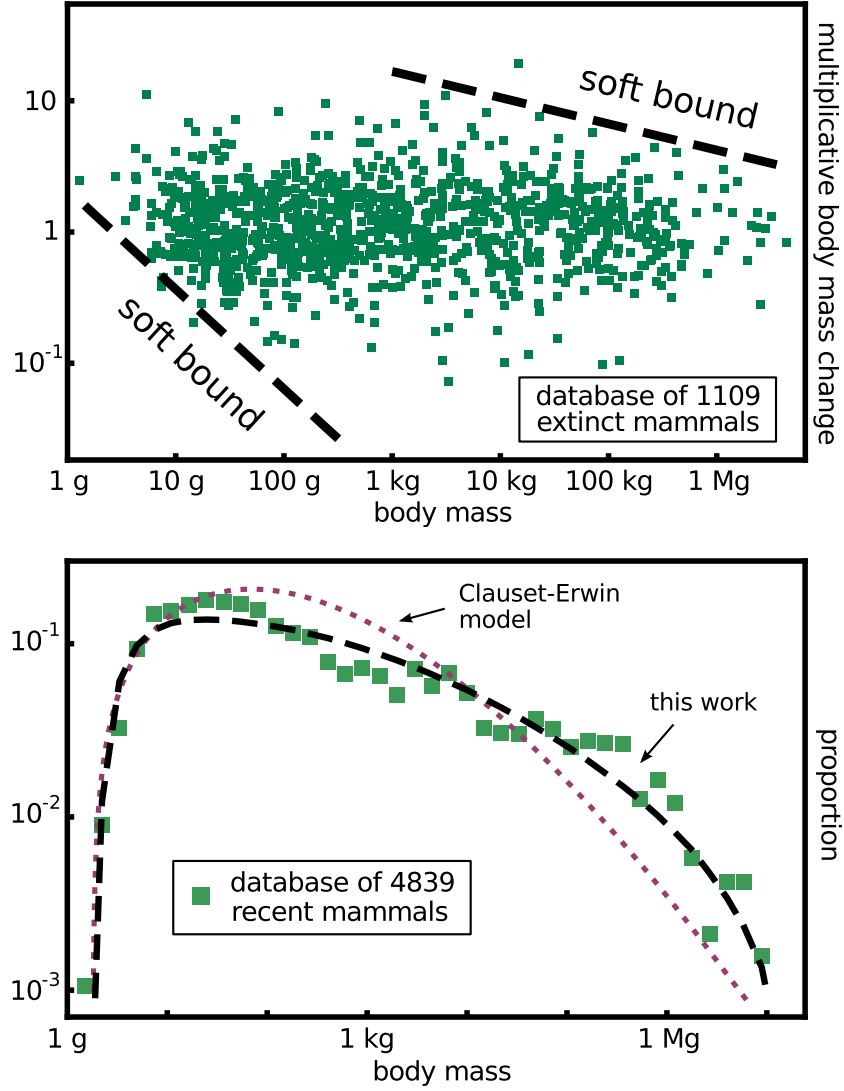


Figure 3: **Application of the bounded diffusion framework to mammalian body-mass data.** (Top panel) Multiplicative changes in body mass for 1109 mammalian species, plotted as a function of ancestor's body mass (squares); data are compatible with the existence of soft bounds (dashed lines), but do not allow to define them. (Bottom panel) The distribution of mammalian body masses is well reproduced by the model (dashed line), and some features appear to be improved with respect to the Clauset-Erwin model (dotted line). Note that the dotted line corresponds to a stationary state, while the dashed line does not.